# Multi-year evaluations of a cloud model using ARM data

# (submitted to J. of Atmos. Sci., Sept. 2008)

PETER W. HENDERSON*

ROBERT PINCUS

Cooperative Institute for Research in Environmental Sciences,

University of Colorado/NOAA Earth System Research Laboratory,

Physical Sciences Division.

* *Corresponding author address:* Cooperative Institute for Research in Environmental Sciences, University of Colorado/NOAA Earth System Research Laboratory, Physical Sciences Division, 325 Broadway Boulder, CO 80305-3337.

E-mail: peterh@colorado.edu

ABSTRACT

The advent of multi-scale models, wherein cloud system resolving models (CSRM) is embedded in each large-scale grid column a model, make increased demands of the CSRM. While higher resolution CSRMs outperform parametric models in case studies, little is known about the performance of more coarse versions, used to reduce computational cost, over a wide-range of atmospheric states.

We exploit long-term lidar and radar retrievals of the vertical structure of cloud at the ARM program's SGP site to evaluate cloud occurrence in 3-year runs of three CSRM configurations of varying resolutions and sophistications. To make the modeled and observed fields more comparable, we use the definition of observed cloud occurrence, based on instrument sensitivity to define cloud in the model. We apply probabilistic measures from ensemble forecast verification that do not require any temporal averaging of the observations, as well traditional performance measures that assume ergodicity.

When thermodynamics is constrained, the bias is relatively small in all runs, suggesting that cloud occurrence is relatively well calibrated in all model configurations. The Brier scores attained by all configurations also suggest considerable model skill. Greater differences in performance are found between seasons than between model configurations during the same season, despite substantial differences between the computational costs of the configurations. Several significant seasonal dependencies are identified, most notably: greater conditional bias, but better timing, of boundary-layer cloud in winter, and substantially less condi-

tional bias in high cloud during summer.

# 1. Why routinely evaluate a cloud model?

Multi-scale models of the atmosphere replace traditional bulk cloud parameterizations in a global model with calculations made by a cloud system resolving model (CSRM) running in each large-scale grid column. This approach is motivated by the repeated ability of CSRMs to outperform simple parametric models in a range of case studies (Randall et al. 2003). However, the CSRMs embedded in multi-scale models are typically used at much lower resolutions than would be used in stand-alone comparisions, and model skill at these resolutions is not a forgone conclusion. More generally, suggestions that CSRMs can be used to link global models and observations (Randall et al. 1996) or to provide the basis for parameterizations in these models (Lock et al. 2000) rely on the ability of CSRMs to make accurate predictions. The predictions can only be assessed by confronting them with observations (Randall et al. 2003) but few quantitative standards of CSRM performance exist. How should such encounters be made and by what measures should success be judged?

Most CSRM evaluations focus on model skill during relatively short case-study periods for which detailed observation are available. Because computational cost is not the driving factor, the model may be run at high spatial resolution. When CSRMs are used in global models, however, they are run at much lower resolution and subject to a much wider range of atmospheric conditions, and evaluation of the CSRM should logically follow suit.

What observations may be used to evaluate CSRMs over such a wide range of conditions? One possibility is the ground-based observatories operated by the Atmospheric Radiation Measurement (ARM) program. These sites combine retrievals from upward-pointing active remote-sensing instruments (radars and lidars) to produce long-term, high-frequency records

of the vertical structure of clouds (e.g. Clothiaux et al. 2000). The main difficulty with using these observations is that they are effectively point-like, complicating comparisons with the larger spatial scales of a model domain.

Traditionally, point measurements of a field are compared to model forecasts by invoking the ergodic hypothesis, which asserts that observations averaged over time are equivalent to the spatial mean of the field. Traditional measures of agreement may then be used. This approach relies heavily on identifying optimal averaging scales, since the ergodic hypothesis fails if significant differences exist between the spatial and temporal statistics of cloud in either the model or the observations, making many measures inappropriate.

An alternative is to apply probabilistic techniques developed to verify ensemble forecasts (Jakob et al. 2004). These techniques are well established in numerical weather prediction but have seldom been applied to cloud models. They are conceptually appealing because they bridge the disparities of scales without reducing the information content of the observations or relying on time averaging. However, it is not clear if this approach has any demonstrable advantage in practice, or even if the measures are sensitive enough to distinguish between the performances of different models.

Here we use both traditional and probabilistic methods to evaluate the performance of a CSRM in predicting cloud occurrence under a wide range of atmospheric conditions. We consider three model configurations of varying resolution and sophistication to quantify trade-offs between model skill and computational cost. We focus on forecasts of cloud occurrence as opposed to continuous fields (e.g. liquid water content) in order to reduce observational uncertainties. Even so, we take care to map model forecasts to the observations by accounting for instrument sensitivity and other observational articfacts.

Section 2 gives details of the CSRM runs including the nudging used to keep the model near the observed thermodynamic state. Section 3 describes the methods that we use to make modeled and observed cloud data more comparable. Section 4 compares various measures used to evaluate model performance; these methods are applied to model predictions using three configurations in section 5. Section 6 discusses the results and offers some possible explanations for the main findings.

## 2. Long model runs

We evaluate the performance of the System for Atmospheric Modeling (SAM) at ARM's Southern Great Plains site. This is the CSRM at the heart of the super-CAM (Khairoutdinov et al. 2005) which is the multi-scale form of the Community Atmosphere Model. We use the term Multi-scale Modeling Framework (MMF) to refer to this type of system.

Our runs are all 3 years long in order to sample as much of the model's probability distribution as possible. Forcing data developed using variational analysis (Zhang et al. 2001; Xie et al. 2004) enable us to drive the CSRM runs over the period 1999–2001 with surface fluxes of latent and sensible heat, large-scale advective tendencies of temperature and moisture, and horizontal winds.

We use 3 configurations of SAM: (i) a standard configuration, like that used in the MMF, namely, a 32-column, 2D domain, oriented east-west with $\Delta x = 4$ km and 28 vertical levels; (ii) a much higher spatial resolution ($\Delta x = \Delta y = 500$ m ) 128 x 128 column, 3D domain with 64 levels, which is capable of better-resolved dynamics; (iii) a configuration with the same low spatial resolution of the standard model, that includes an Intermediate Prognostic High

Order Closure (IPHOC) of turbulence, to improve the representation of shallow cumulus and its transition to deep convection (Cheng and Xu 2008). IPHOC treats subgrid-scale transport and, in particular, allows for sub-grid scale fractional cloudiness.

All configurations have cyclic boundary conditions and use a stretched vertical grid, such that the spacing between levels increases with height, and is typically 100–500 m in the planetary boundary layer (PBL). The domain extends to $\sim$ 28 km, and Newtonian damping is applied to the upper 1/3 of the model domain to suppress gravity waves (Khairoutdinov and Randall 2003).

Instantaneous model output is collected hourly, producing $\sim$ 2.6 x $10^4$ values at each model level in each column of the domain. At this time-lag, observed cloud occurrence has autocorrelations of between 0.5–0.7, so not much information is gained by sampling more frequently.

Our goal is to assess the model's ability to predict cloud occurrence given an observed thermodynamic state. However, even when forced with observed large-scale fields, significant biases in thermodynamic fields appear in a matter of weeks (Khairoutdinov and Randall 2003). To reduce this bias, temperature is nudged towards the observed soundings. Water vapor is not nudged, because it is already heavily constrained by surface precipitation. Our selection of a nudging time scale for temperature $\tau_T$ is based on 15 long runs covering a wide range of $\tau_T$. In the model's heat equation, the nudging term $(T_{mod} - T_{obs})$ / $\tau_T$ is non-physical, so its relative magnitude must be kept as small as possible compared to the advective tendency (Ghan et al. 1999). After balancing this need with the magnitudes of concomitant errors in thermodynamics and surface precipitation, we choose $\tau_T = 24$ hours as a default value. Since there is no pressure gradient term in the CSRM's prognostic equations,

the winds are nudged on a relatively shorter time-scale; a period of 2 hours is used, in keeping with previous authors (e.g. Khairoutdinov and Randall 2003).

# 3. Reducing the gap between models and observations

*a. Mapping model cloud to observed cloud*

We use instantaneous observations of (binary) cloud occurrence as our evaluation data; this cloud mask is obtained from the ARM program's Active Remotely-Sensed Cloud Locations (ARSCL) lidar and radar cloud-boundaries product (Clothiaux et al. 2000), hereafter referred to simply as ARSCL. The observations are made on a finer grid than that of the model; we verify forecasts using those observations closest to the model levels.

ARSCL's vertical grid is defined at 512 altitudes, starting at 105 m, with a spacing of 45 m, spanning an atmospheric column of $\sim$ 23 km, but the maximum altitude of the large scale forcing data is 100 hPa ($\sim$ 17 km), so we ignore all values above this.

To make the definition of modeled and observed cloud consistent, we take the definition of cloud used in the observations and map this to the model — in the spirit of ISCCP instrument simulation (Klein and Jakob 1999). This requires us to replicate ARSCL's algorithm, which merges measurements from vertical pointing, narrow-beam millimeter cloud radar, micropulse lidar. We construct separate radar and lidar cloud masks, which are then combined to make a final mask for each column of the model's domain.

### 1) Radar cloud mask

We simulate the reflectivities that ARM's 35 GHz millimeter cloud radar (MMCR) would observe, using the QuickBeam radar simulator (Haynes et al. 2007), which accounts for attenuation of the radar beam by both atmospheric gases and hydrometeors in its calculations of the ground-based reflectivity profiles. For every column of the model's domain the hourly, instantaneous mixing ratio profiles of each of its hydrometeor species are input to the simulator, together with those of temperature and relative humidity.

All condensed species are assumed to be spherical with density dependent on diameter only. We assume exponential particle size distributions for cloud ice, rain, snow and graupel; a log-nomoral distribution is assumed for cloud liquid. To speed up the simulation, Mie calculations are approximated by lookup tables — the reflectivity errors induced by doing this are typically less than 2 dB (Haynes et al. 2007).

The resultant reflectivities are compared to a detection threshold, $\mathrm{dB}_{lim}$ which varies with height (P. Kollias, personal communication) to account for attenuation of the beam by gases and geometric spreading. Fig.1 shows the p.d.f. of non-zero simulated reflectivites as a function of height. Because most values are substantially greater than the threshold, our results are not particularly sensitive to our choice of $\mathrm{dB}_{lim}$.

### 2) Lidar cloud mask

ARSCL is primarially based on radar observations but uses optical lidar retrievals to detect thin clouds, and to identify precipitation falling below clouds, aiding the determination of cloud base during such events (Clothiaux et al. 2000). Since we are only interested in the

altitude at which the optical beam becomes extinguished by cloud (rather than simulating reflectivites) we proceed by approximating the extinction $k_{ext}$ of the lidar beam.

The CAM is used as a parent model in the MMF, so we use the effective radii $r_e$ assumed by CAM for radiation calculations to infer droplet number density $N$ from local cloud water and cloud ice concentrations. Assuming a scattering efficiency of 2 gives $k_{ext} \approx 2\pi r_e^2 N$, and we define cloud occurrence (according to the lidar) whenever $k_{ext} > 0$. At any given altitude, the beam is assumed to be fully extinguished whenever the optical depth exceeds 2.

### 3) Final cloud mask

The lidar and radar cloud masks are combined to produce the final mask. This takes the value of the lidar mask when and where the lidar is known not to have been attenuated, otherwise the value of the radar mask is used; and this completes our replication of the ARSCL algorithm.

The spatial mean of the final cloud mask is taken at each model level, hourly; this mean is the probability of cloud $p$ which we evaluate against the dichotomous observations.

### 4) Subgrid-scale sampling

The IPHOC scheme used in one of our runs allows for condensation to occur in only a fraction, PCON of each model grid cell. We account for this possibility using a Monte Carlo sampling technique similar to those used to represent sub-grid scale structure in global models (e.g. Klein and Jakob 1999; Räisänen et al. 2004). For each SAM model column we construct a single sample column, ensuring that the mean of many sample columns reproduces the

cell-by-cell values of condensate amount and of PCON. Columns are constructed using the maximum overlap assumption, though results using random overlap are essentially the same.

# 4. Measures of model performance

In forecast verification, the practice of using only one or two parameters of the univariate distribution of model errors is referred to as measures oriented and that based on the joint distribution of the forecasts and observations is referred to as distributions oriented. We use a mixture of both and demonstrate a link between them.

Performance metrics can sometimes be broken down into components estimating different aspects of a model's predictive ability. These are referred as attributes (Murphy 1993) and in this work we make use of four, as follows. Bias is the correspondence between the mean forecast and mean observation. Reliability (REL) is the correspondence between the conditional mean observation and the conditioning predictions. Resolution (RES) is the ability to resolve observed events into subsets with characteristically different outcomes. Uncertainty (UNC) is the variance of the observations, which is independent of the model; large values make skillful forecasts difficult.

The resolutions of short-term weather forecast models are typically an order of magnitude less than their reliabilities (Stanski et al. 1989) — unlike reliability, resolution cannot be improved by calibration techniques (Atger 2003) and therefore provides an invariant measure of a model's ability.

*a. Mean squared error and Brier's probability score*

A traditional performance measure of a model's performance in predicting a continuous variable $x$ is the mean squared error (MSE). This can be broken down into a bias and random error $\text{MSE} = \text{var}(p - x) + (\bar{p} - \bar{x})^2$, where $p$ denotes prediction, over-bar denotes temporal mean and var is the variance.

To calculate the MSE for a variable where observations $o$ are 1 if the event occurs, and 0 if it does not, requires us to transform the observations into a continuous field by averaging them over some time-period of length $L$. However, if $p$ represents an instantaneous spatial field, then taking the MSE only makes sense if the temporal and spatial statistics are approximately equivalent, and even if this is true, we must still make an appropriate choice of $L$.

Alternatively, we can use the original (dichotomous) observations and calculate the Brier score $b$, a well established measure used in the verification of operational numerical weather prediction models, which removes the need to average over time. Though they are often discussed separately, $b$ can be defined as a limiting case of the MSE. As we decrease the averaging period, so that $L \to 1$, we find that $x_i \to o_i$, the $i^{th}$ instantaneous observation, i.e.

$$\lim_{L \to 1}(\text{MSE}) = \frac{1}{N}\sum_{i}^{N}(p_i - o_i)^2 = b, \tag{1}$$

where the $p_i$ are probabilities of the original dichotomous event occurring.

In this work we exploit a decomposition of $b$ into components measuring key attributes of model performance, namely: REL - RES + UNC. More formally, by dividing the probability range $[0, 1]$ into $K$ probability classes (bins) we can write

$$b = \frac{1}{N}\sum_{k}^{K} n_k\Big((p_k - \bar{o}_k)^2 - (\bar{o}_k - \bar{o})^2\Big) + \bar{o}(1 - \bar{o}), \tag{2}$$

where $\bar{o}_k$ is the mean observed frequency of occurrence for the $k^{th}$ class, containing $n_k$ events, and $\bar{o}$ is the observed climatology; other decompositions are also possible (Murphy 1996). The best possible value is $b = 0$, wherein REL $= 0$ and RES $=$ UNC. However, this perfect score is only attainable by a model predicting (correct) extreme probabilities of 0 or 1, making the predictions deterministic; we can think of this as the asymptotic limit of a probabilistic model making increasingly resolved predictions.

Values of $b$ typically lie in the range [0.10, 0.25] for numerical weather model forecasts and scores $> 0.3$ (in most cases) represent poor predictions; scores for forecasts of rare events tend to be better and will usually be $< 0.10$ (Stanski et al. 1989). High skill is implied by $b <$ UNC.

*b. What do the performance measures tell us?*

The multiple runs made to determine $\tau_T$ provide us with an opportunity to investigate how different performance measures vary with this parameter, results of which are shown in Fig.2. As well as looking at their relative sensitivities, we establish the ranges spanned by the scores of these very different runs, since we can compare these to the ranges for other sets of runs, such as those covering different model configurations.

In this case the similarity between the MSE of the temporal cloud-fraction and $b$ is very evident; both scores have almost identical variation with $z$ and $\tau_T$. This means that averaging the observations over time produces similar results to averaging over the model's domain, which suggests that the ergodic hypothesis holds here. In turn, this validates our use of the MSE and its components here.

Above 9 km, the mean bias grows rapidly with $\tau_T$, and correlates with a monotonically increasing negative model bias in $T$ (not shown), although the MSE remains dominated by its random component $\sigma^2$. The mean bias, REL and $\sigma$ increase with $\tau_T$, but timing, as indicated by RES, is similarly poor for all $\tau_T$, suggesting that it contains information that the others do not. At lower altitudes RES decreases steadily with $\tau_T$ and demonstrates a larger range than REL.

Reliance upon any single measure is limited, since (alone) they tell us nothing about the circumstances in which particular types of errors occur. We can go a stage further by looking at REL and RES as a function of probability, and the relative contribution that each probability range makes to each of these components; this approach will be taken up in the next section.

# 5. Model evaluations

Here, we compare the performances of the 2D, 2D+IPHOC and 3D model configurations. The 3D run is approximately a factor of 1000 more computationally expensive than the standard 2D run, and the IPHOC run around a factor of four. One might reasonably imagine the scores of these configurations to reflect this. Fig.3 shows the aggregate scores and Fig.4 shows those for events restricted to the periods April–September and October–March, which we refer to as summer and winter, respectively.

*a. Aggregate scores*

The long term performance of all three configurations is similar, as measured by any score (see Fig.3) . The mean bias is low in all runs, suggesting reasonable overall calibration of cloud; consequently, almost all of the MSE consists of random error. At many altitudes, the Brier scores attained by all configurations suggest considerable skill, with better performance below 8 km where $b <$ UNC.

In all configurations, the greatest conditional biases (REL) are seen for high cloud, particularly so in the IPHOC and 3D runs between 10–13 km. The 3D run is the most conditionally biased in the PBL, but values here are typically a factor of 3–4 less than for high cloud, in all runs. Performance in timing (RES) has a pronounced maxima in the PBL and drops off rapidly above 9 km.

*b. Seasonal scores*

For all model configurations, the performance during summer and winter is markedly different. The relative differences between the runs are also seasonally dependent (see Fig.4). In fact, differences in skill between seasons are typically greater than differences between configurations during the same season. Given the range of computational costs of the configurations, this is surprising.

Model errors are typically greater in winter than in sumer; the most notable exception to this being worse timing below 9 km in summer. At some altitudes (e.g. PBL for the 3D run ) the overall bias is comprised of opposing seasonal biases.

We now look in more detail at conditional bias and timing as a function of $p$ close to 1

km and 10 km. These altitudes are of interest for a number of reasons: (i) they are close to the maxima in observed cloud; (ii) large seasonal differences in performance occur here; (iii) relatively large differences exist here between the scores of different model configurations; and (iv) the variance of observed cloud (UNC) is approximately the same for boundary layer and high cloud, which means that the Brier scores at these altitudes are directly comparable.

We construct attributes diagrams for each altitude and season (Fig.5) by conditionally sampling the observations using the forecasts of cloud fraction (i.e. probability of cloud, $p$). The diagrams plot observed frequencies of cloud occurrence against the corresponding forecast probabilities, as well as forecast distributions and observed climatologies. The diagrams are augmented with information about the relative contributions that each probability range makes to the conditional bias (ideally zero) and ability in timing (ideally large).

### 1) Scores & contributions from forecast populations

Extra wintertime conditional bias in PBL and high cloud are two of the main seasonal differences identified by the scores, here we investigate which parts of the forecast distribution are responsible. For high cloud, all configurations over-predict most $p$ (indicated by values lying below the 1:1 line of the attributes diagram) particularly in winter. In the PBL, all configurations over-predict mid–high $p$, but also under-predict low cloud-fractions (lie above the 1:1 line) particularly in summer.

In summer, high cloud is most conditionally biased in the 3D run. This is due to extra over-forcasting of mid-high $p$ and greater population of this probability range, as opposed to extra over-forecasting of $p = 1$ predictions, which occur more often in the other runs. Even

though slightly more cloud is observed at 10 km in the summer, in all runs the number of $p = 1$ predictions at this altitude is an order of magnitude less than in winter.

In all runs, high cloud is substantially more conditionally biased in winter than in summer. No rapid increase in over-prediction occurs for high $p$ (a distinctive feature of the summertime attributes diagram) but the contributions from this range are responsible for the seasonal differences. The greatest contributions are for $p = 1$ predictions, because these are more numerous — IPHOC forecasts most and is therefore most biased here. Although we do not show the attributes diagram for 7 km, the 3D run performs best here, and IPHOC worst (see Fig.4) because of the extra number of $p = 1$ predictions in the former, since the observed frequencies corresponding to these predictions are similar in both runs.

For any given season, the timing of cloud is remarkably similar in all runs; the smallest seasonal differences are at 10 km. Here, most contributions to good timing (high RES) come from forecasts of clear sky, intermediate cloud-fractions in summer, and 100% cloud fractions in winter.

The p.d.f.s of wintertime forecasts in the PBL are similar to those of high cloud, with the conditional bias again being determined by the number of high $p$ predictions; however, here the 3D run is worst, because it has the greatest number of $p = 1$ forecasts. In the PBL, the greatest number of clear and least mid–high $p$ forecasts occur during summer which consequently demonstrates the least conditional bias in all configurations.

Generally, more seasonal differences in timing are seen in the PBL than aloft, with greater contributions being made by clear and over cast ($p = 0, 1$) forecasts in winter than in summer. This gives rise to the better timing of wintertime cloud here, most of which is attributable to the increased number of predictions of high cloud-fraction. Furthermore, this result can

be extended to explain the worse timing of most cloud below 9 km in summer.

# 6. Discussion & conclusions

*a. Main findings*

Our evaluation demonstrates that greater differences in performance occur between seasons than between model configurations during the same season, despite substantial differences between the computational costs of the configurations. Several significant seasonal dependencies have been identified, most notably greater conditional bias but better timing of PBL cloud in winter and substantially less conditional bias in high cloud during summer.

The mean bias is relatively small for all three model configurations which suggests that cloud occurrence is relatively well calibrated. The Brier scores attained by all configurations also suggest considerable model skill at many altitudes, when compared to those typically achieved by numerical weather models. We conclude that this model demonstrates skill in predicting cloud occurrence, given the proper thermodynamic state, over a wide range of atmospheric conditions.

The similar behavior of the MSE and $b$ for cloud-occurrence suggests that the ergodic hypothesis holds for these simulations. This means that we can approximate the first few statistical moments of the instantaneous, spatial cloud-fraction with point-like, temporal cloud-fraction, and vice versa, making application of the traditional measures valid.

A relatively large range of values is found between the scores for different seasons, often greater than the range seen between runs using very different $\tau_T$. This suggests that the

scores (particular REL and RES) are sufficiently sensitive to be able to identify differences in model performance. This increases our confidence in interpreting similar scores as real similarity in performance, rather than a lack of precision of the measures.

Might short-term model performance change if we stop constraining temperature? Cloud fields from the long runs could be used to initialize short free-running forecasts at regular intervals throughout the 3-year period. Performance of these non-nudged runs can then be compared to those of the nudged runs.

At and above 10 km, all three model configurations produce more wintertime cloud than is observed, and they are all substantially more conditionally biased than in summer; the same is also true in the PBL. Furthermore, the timing of PBL cloud and its dramatic seasonal dependence are almost the same in all configurations. The similarities of the scores of all configurations suggests that conditional bias in cloud is more to do with the CSRM (SAM) than the details of a particular configuration.

All seasonal dependencies are strongly influenced by contributions from the predictions of high cloud-fractions; increased numbers of these predictions ($n_k$ in Eq.(2)) increase the weighting of this part of the forecast distribution, which often dominates the score, giving rise to greater conditional bias. The timing of cloud in the PBL is a partial exception to this pattern, since it is as influenced by the number of clear-sky predictions; the timing of summertime cloud also has more contribution from low–mid values.

During winter, there is more stratiform cloud in the PBL and large-scale frontal cloud-systems; whereas in summer, there is more small-scale shallow convection, deep convection and anvil cirrus. The local and more intermittent nature of summertime cloud may be the main reason for worse timing in the summer season. The contribution that clear-sky

predictions make to the timing of low cloud is greater in winter than in summer because the mean observed cloud fraction is higher in this season. Conditional sampling on variables relevant to the predominant cloud-type may shed light upon the mechanisms responsible.

*b. Similarity of inter-configurational scores*

Despite significant differences in their sophistication, spatial resolution and computational overheads, the scores for each configuration are remarkably similar. No significant disadvantages have been found in restricting the CSRM to its standard 2D configuration; this is consistent with the findings of (Khairoutdinov and Randall 2003) based on runs (of up to one month in length) covering numerous 2D and 3D configurations of an earlier version of this model. While this is favorable for use in the MMF, it also suggests that the cloud model's deficiencies are deeper than can be ameliorated by simple changes.

One possible reason for the similarity of scores for different model configurations is that ice microphysics is loosely constrained in SAM. Aspects of this (e.g. ice fall-speed) could be responsible for a significant amount of the conditional bias, random error and timing seen in all runs.

Some similarity in performance could also be attributable the effects of nudging. However, low correlations ($< 0.2$) between point-wise thermodynamic and cloud errors (not shown) suggests that cloud errors are not significantly influenced by nudging temperature.

We know that there are errors in the the large scale forcing fields and soundings. It is possible that these errors have such a large influence on thermodynamics that they reduce scores enough to hide comparatively smaller inter-configurational differences. In particularly:

errors in the advective tendencies, which are less accurate than the soundings; area-averaged surface precipitation which has the most influence on the advective tendencies; and spatial scale aliasing in fields with large subgrid-scale variability, such as water vapor and winds, and during severe weather. Multiple physically consistent forcing data could be created, spanning the range of uncertainty in the observed fields Zhang et al. (2001). These could be used to drive different members of ensemble runs of the CSRM to explore sensitivities to observational uncertainty (Hume and Jakob 2007).

It is also possible that all configurations show approximate equal skill because we are looking only at (binary) cloud occurrence and that differences may exist in the structure of the (continuous) hydrometeor fields. However, mean total condensate is very similar in all three cases (not shown) suggesting that this is not the case.

The inclusion of IPHOC has not improved the model's performance in predicting cloud occurrence. IPHOC generally makes the development marine boundary layer cloud and shallow convection smoother in time (Cheng and Xu 2008), but these advantages do not translate to continental cloud at the horizontal and vertical spatial resolutions considered here. Further study, focussing on the PBL and using alternative grid spacings may provide further insight.

*c. Differences between inter-configurational scores*

While the distributions of cloud fraction predicted by each of the model configurations are similar in each season, some differences are found, particularly for mid-high $p$. Contributions from this range (especially $p = 1$) often dominate the performance scores, such that the configuration with the greatest population of this range, typically has the greatest conditional

bias.

Although very little cloud is observed above 10 km in winter, more occurs in summer, and it is interesting that the greatest inter-configurational differences in conditional bias are found here, such that more sophisticated configurations are most biased. Inspection of the attributes diagrams for these altitudes (not shown) confirms that this is due to the extra number of predictions of 100% cloud fraction, compared to those of the standard 2D configuration — more such predictions are actually made in winter, however, more equally so in all runs. Identification of the circumstances under which the 3D and IPHOC runs produce extra, very high cloud may provide further insight.

# 7. Acknowledgments

We are grateful to Marat Khairoutdinov for providing and supporting SAM; Anning Cheng and Kuan-Man Xu for implementing IPHOC; NOAA-HPCS for use of Jet and Christopher Harrop for advice; Roger Marchand for help with the radar-simulator and Patrick Hofmann for helping build a QuickBeam interface; and Eugene Clothiaux, Christian Jacob and Tom Hamill for discussion and insight. This research was supported by the Office of Science (BER), U.S. Department of Energy, under grants DE-FG03-06ER64182 and DE-FG02-03ER6356.

# REFERENCES


Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Weather Rev.*, **131 (8)**, 1509–1523.

Cheng, A. and K.-M. Xu, 2008: Simulation of boundary-layer cumulus and stratocumulus clouds using a cloud-resolving model with low- and third-order turbulence closures. *J. Meteor. Soc. Japan*, **(in press)**.

Clothiaux, E. et al., 2000: Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART sites. *J. Appl. Meteorol.*, **39 (5)**, 645–665.

Ghan, S. J., L. R. Leung, and J. McCaa, 1999: A comparison of three different modeling strategies for evaluating cloud and radiation parameterizations. *Mon. Weather Rev.*, **127 (9)**, 1967–1984.

Haynes, J., R. Marchand, Z. L. A. Bodas-Salcedo, and G. Stephens, 2007: A multi-purpose radar simulation package: Quickbeam. *Bull. Am. Meteorol. Soc.*, **88**, 1723–172.

Hume, T. and C. Jakob, 2007: Ensemble single column model validation in the Tropical Western Pacific. *J. Geophys. Res.*, **112 (D10)**, D10 206.

Jakob, C., R. Pincus, C. Hannay, and K.-M. Xu, 2004: Use of cloud radar observations for model evaluation: A probabilistic approach. *J. Geophys. Res.*, **109 (D03202)**.

Khairoutdinov, M. and D. Randall, 2003: Cloud resolving modeling of the ARM summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities. *J. Atmos. Sci.*

Khairoutdinov, M., D. Randall, and C. DeMott, 2005: Simulations of the atmospheric general circulation using a Cloud-Resolving Model as a superparameterization of physical processes. *J. Atmos. Sci.*, **62**, 2136–154.

Klein, S. A. and C. Jakob, 1999: Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Weather Rev.*, **127 (10)**, 2514–2531.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Mon. Weather Rev.*, **128**, 3187–3199.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281–293.

Murphy, A. H., 1996: General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Weather and Forcasting*, **124**, 2353–2369.

Räisänen, P., H. W. Barker, M. F. Khairoutdinov, J. N. Li, and D. A. Randall, 2004: Stochastic generation of subgrid-scale cloudy columns for large-scale models. *Q. J. R. Meteorol. Soc.*, **130 (601)**, 2047–2067.

Randall, D. et al., 2003: Confronting models with data: The GEWEX cloud system study. *Bull. Am. Meteorol. Soc.*, **84**, 455–469.

Randall, D. A., K.-M. Xu, R. J. C. Somerville, and S. Iacobellis, 1996: Single-column

models and cloud ensemble models as links between observations and climate models. *J. of Climate*, **9**, 1683–1697.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: A survey of common verification methods in meteorology. Tech. Rep. 8, 358, WMO.

Xie, S., R. T. Cederwall, and M. Zhang, 2004: Developing long-term single-column model/cloud system-resolving model forcing data using numerical weather prediction product constrained by surface and top of the atmosphere observations. *J. Geophys. Res.*, **109 (D01104)**.

Zhang, M. H., J. L. Lin, R. T. Cederwall, J. J. Yio, and S. C. Xie, 2001: Objective analysis of ARM IOP data: Method and sensitivity. *Mon. Weather Rev.*, **129**, 295–311.

# List of Figures

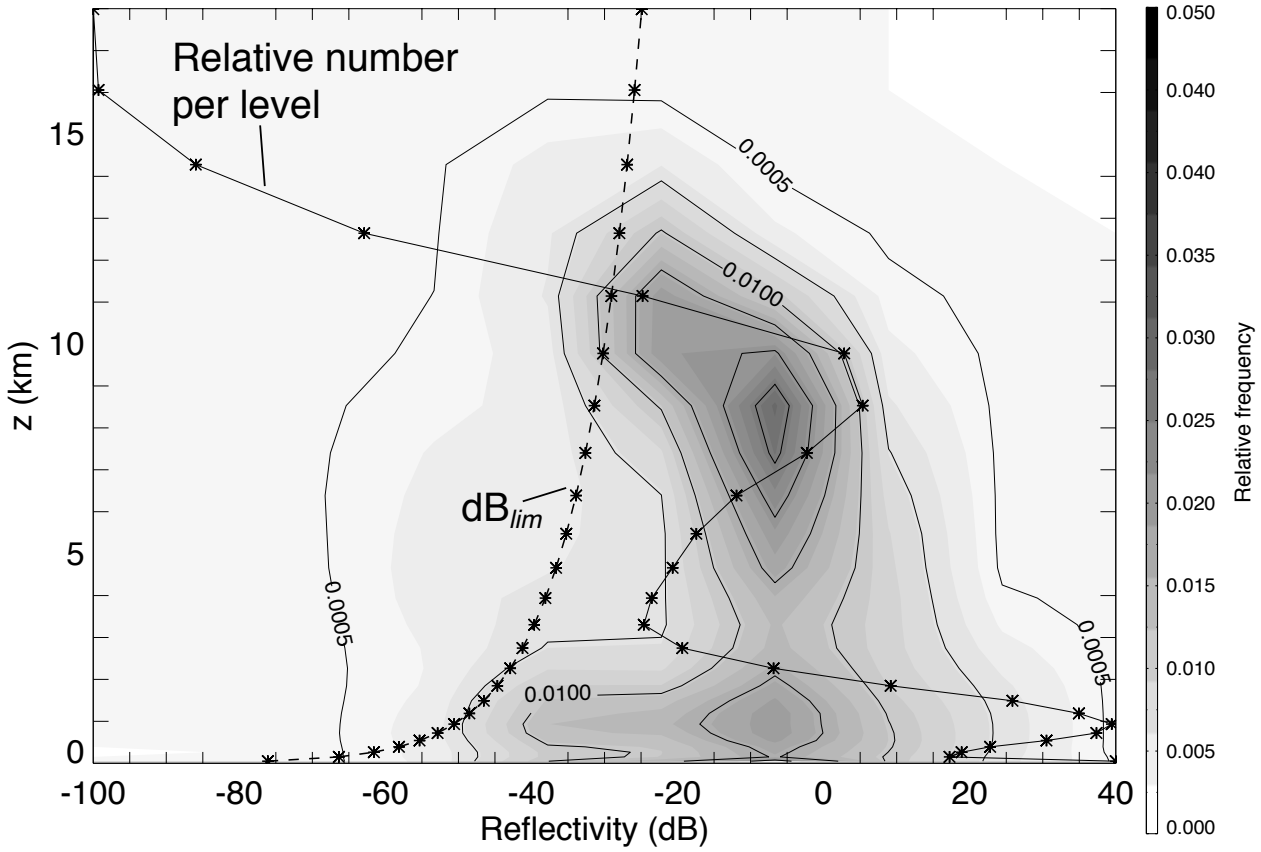


FIG. 1. Normalized distribution of radar reflectivites ($> -100$ dB) simulated using the cloud liquid, cloud ice, rain, snow and graupel from all columns of a 2D 3-year run. Most values are greater than the sensitivity threshold $\mathrm{dB}_{lim}$ used to define cloud in the model.

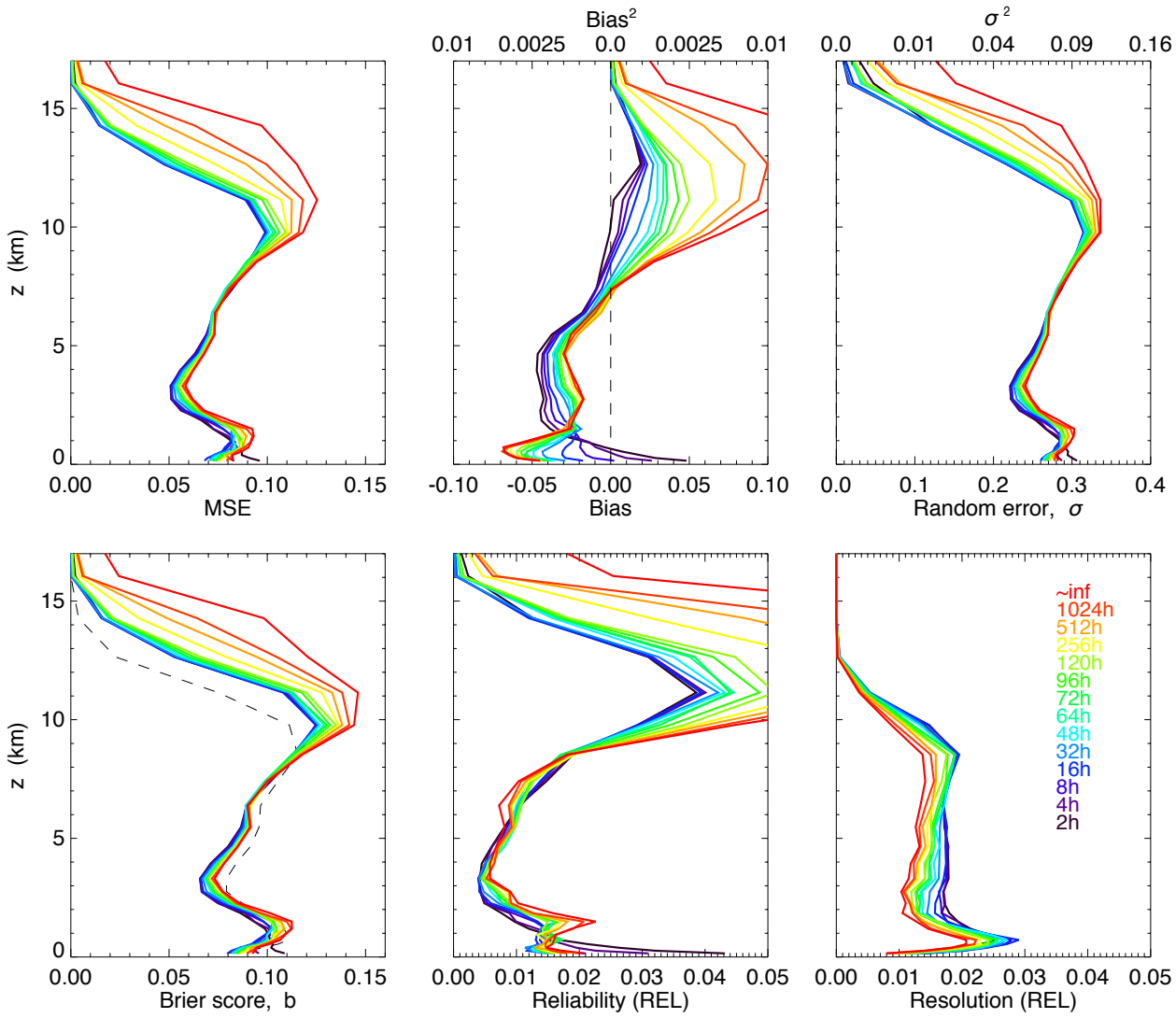


FIG. 2. Performance scores of 14 separate 3-year runs at the SGP, each using different nudging periods, $\tau_T$ (n.b. inf denotes no nudging of $T$). Overall bias and MSE are with respect to observed temporal (hourly) cloud-fraction. All other measures are with respect observed instantaneous cloud occurrence. The dashed line in the plot of Brier score is UNC.

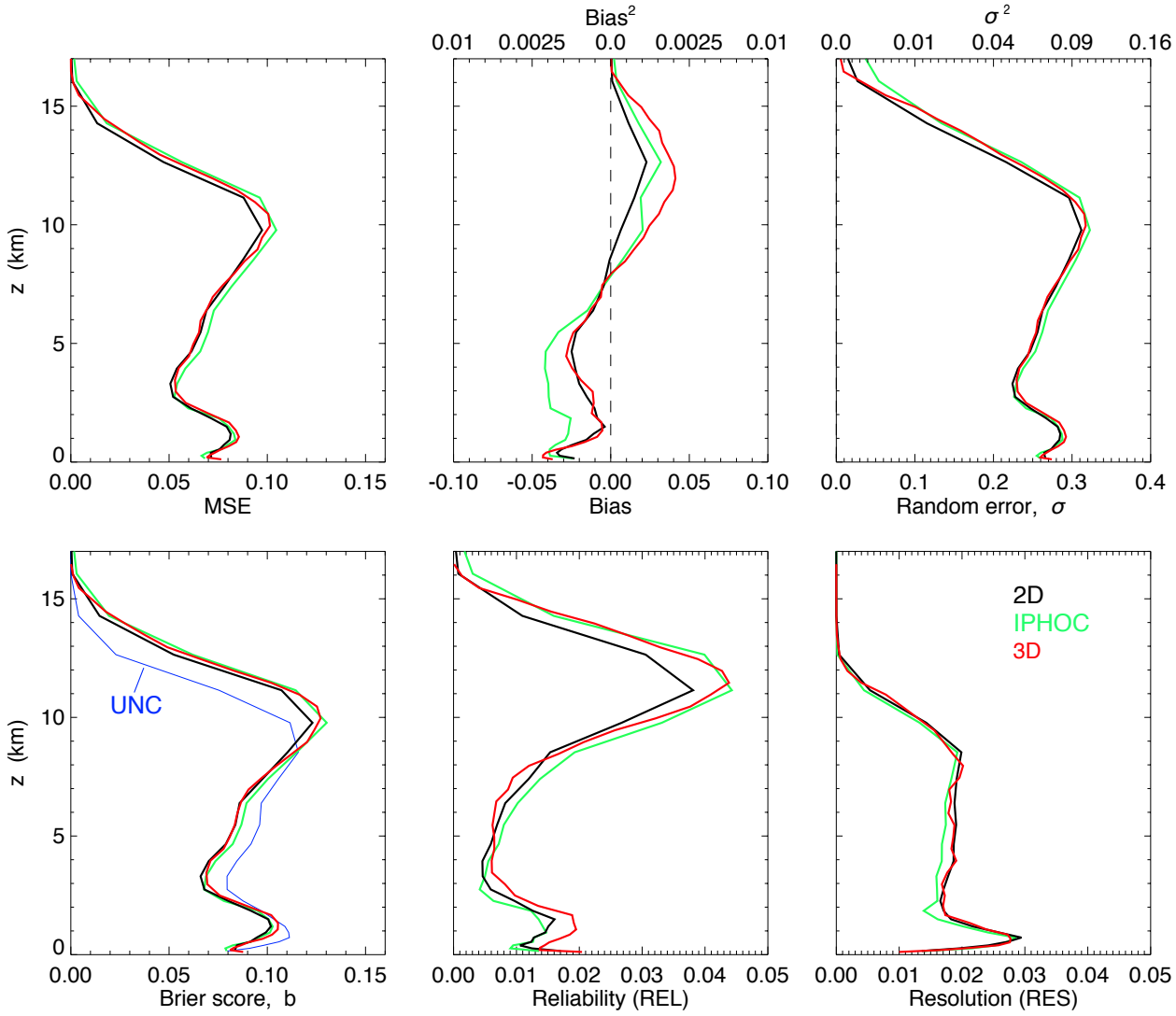


FIG. 3. Performance scores of 3-year runs with different model configurations: (i) standard 2D domain, (ii) 2D + IPHOC, and (iii) higher resolution 3D domain. All runs use $\tau_T = 24$ hours.

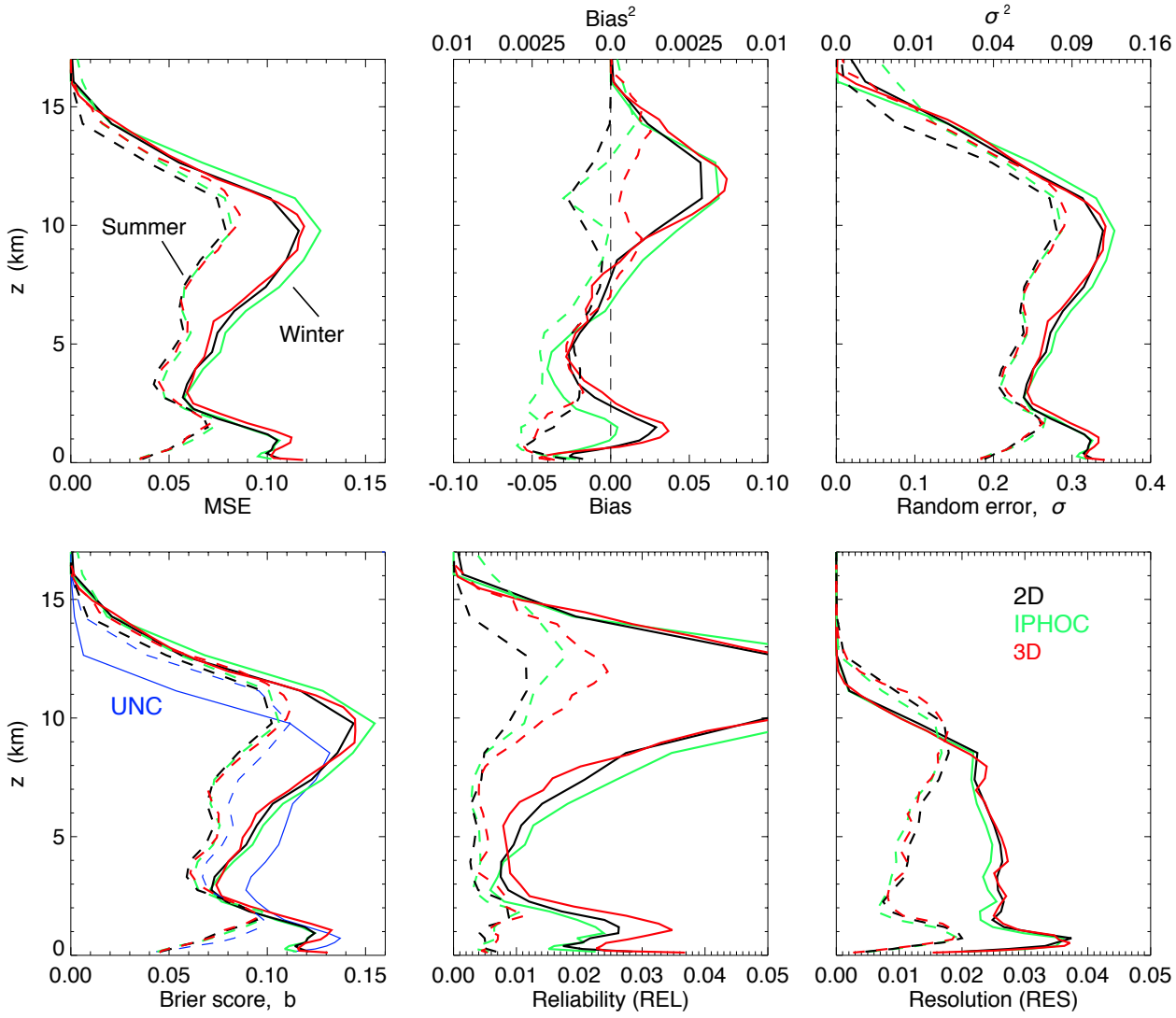


FIG. 4. As for Fig.3, restricted to events from April–September (dashed) covering boreal summer, and events from October–March (solid) covering boreal winter.

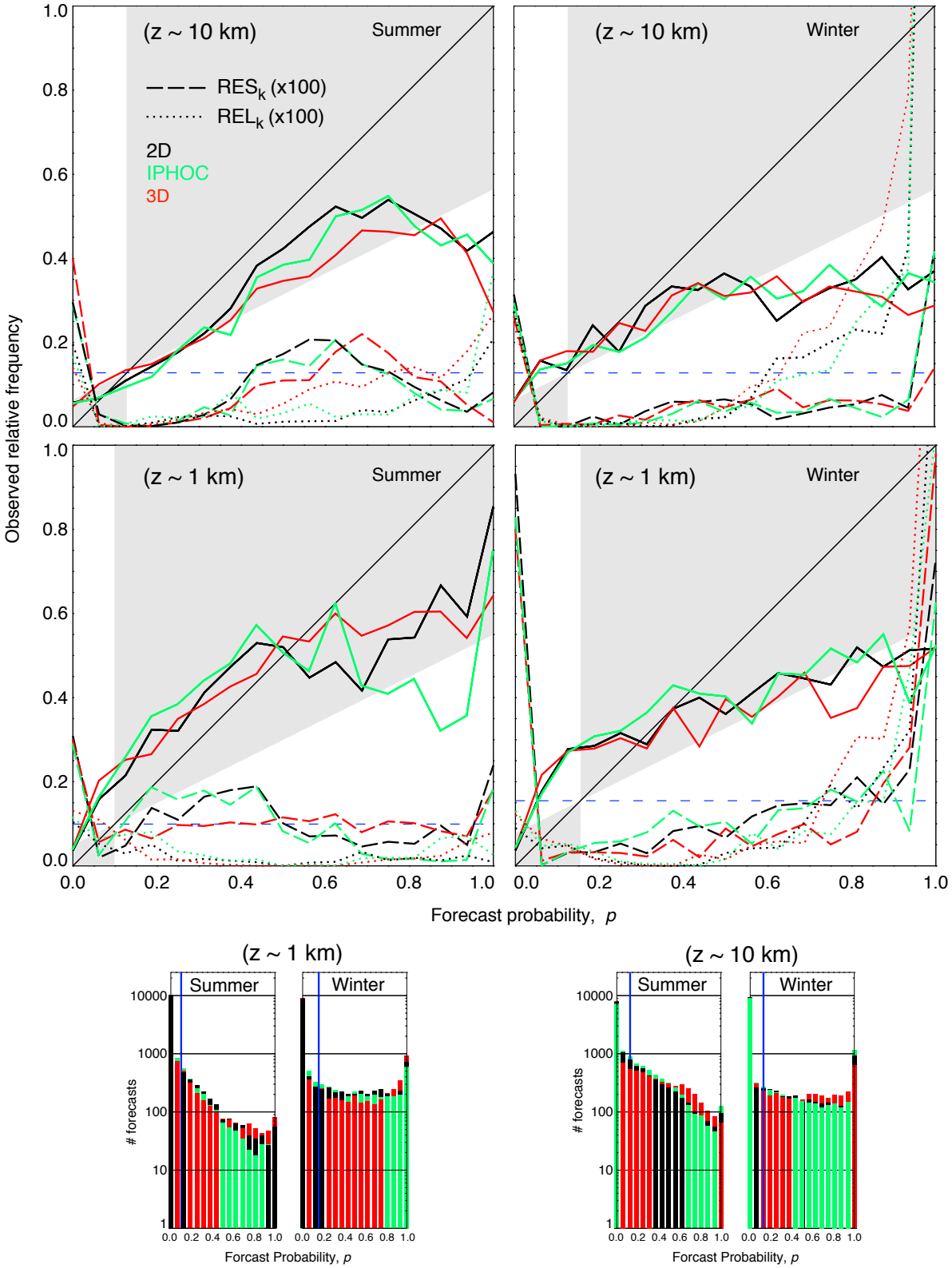




FIG. 5. Augmented attributes diagrams covering boreal summer and winter for high and low cloud. The distance of the solid lines from the 1:1 line indicates reliability and their distance